CHAPTER 1

GENERAL INTRODUCTION

CHAPTER 1

GENERAL INTRODUCTION

THE OCEANS

A world under pressure

Strong anthropogenic pressure rapidly induces changes in the oceans. Oceans are directly exploited by fisheries, gravel and mineral extraction, construction of oil platforms and wind mill farms, the use of wave and tidal energy, installation of pipe lines and distribution cables (Halpern *et al.*, 2008). Moreover, other non-intentional effects such as organic and inorganic pollution, the emergence of invasive species, ocean acidification and climate change operate. As a result, over 40% of the world's oceans are heavily affected by human activities and few, if any areas remain untouched (Halpern *et al.*, 2008) (Fig. 1.1).

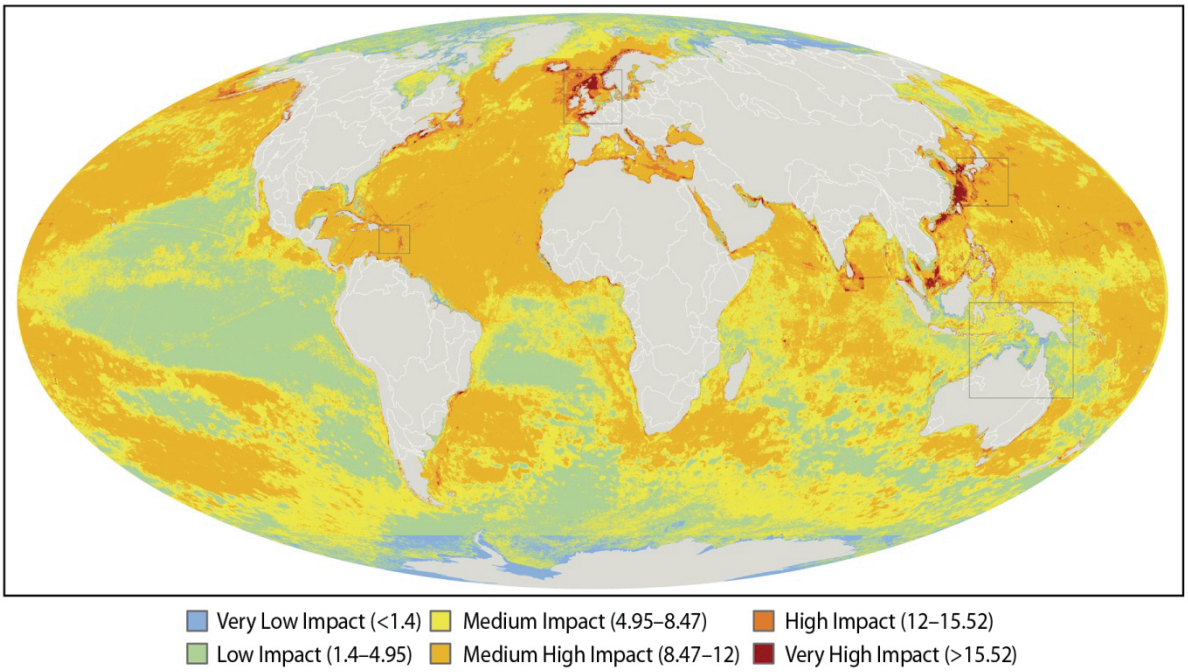

*Fig. 1.1. Global Map of Human Impacts to Marine Ecosystems (*from *Halpern* et al.*, 2008)*

Some of these impacts strongly affect local biodiversity: numerous publications report on the impacts of fisheries (Pauly *et al.*, 1998; EC, 2009; FAO, 2010; Froese *et al.*, 2010; Shephard *et al.*,2010) and climate change related effects (Caldeira and Wickett, 2003; Edwards and Richardson, 2004; Perry *et al.*, 2005). Although marine extinctions are not easily uncovered, it has been shown that regional ecosystems are rapidly losing species and functional groups (Worm *et al.*, 2006).

Global awareness that our ecosystems need protection was raised at the UN Convention of Biological Diversity in 1992 in Rio de Janeiro. The main outcome of this convention consists of two key documents: the Rio declaration and the Agenda 21. One of the objectives in Agenda 21 is to improve the conservation of biological diversity and the sustainable use of our biological resources. A significant reduction of biodiversity loss by 2010 was advocated at the 2002 Convention on Biological Diversity at The Hague (Butchart *et al.*, 2010). This target was further refined in 2006 (CBD, 2006). Specifically, at least 10% of each of the world's ecological regions should be effectively conserved and areas of particular importance to biodiversity should be protected (Toropova *et al.*, 2010). Targets focusing on specific biomes such as the conservation of marine and coastal areas were specified. Marine Protected Areas (MPA's) should be established to protect species, maintain productivity, and preserve nursing grounds for fishes or to protect complete ecosystems (CBD, 2006). A drawback however is that the degree of protection and the activities allowed in these MPA's are not well defined and may vary considerably (Toropova *et al.*, 2010). At a European level, the European Commission issued an action plan to halt the loss of biodiversity by 2012 for the marine environment (EC, 2006). Therefore, the European Commission agreed on two directives to oblige member states to designate MPA's in the frame of the Natura 2000 network:

- The European Birds Directive (2009/147/EC) aims to protect all European wild birds and the habitats of listed species, in particular through the designation of Special Protection Areas (SPA). In relation to the marine environment this is translated in the designation of habitats of sea birds.
- The Habitats Directive (92/43/EEC) focuses on creating a network of Special Areas of Conservation (SAC).

Other European strategies which can be deployed to protect the marine environment are:

- Integrated Coastal Zone Management (2002/413), which promotes the sustainable management of coastal zones, while balancing environmental, economic, social, cultural and recreational objectives all within the natural limits.
- The Water Framework Directive (WFD) (2000/60/EC), which obliges the member states to achieve good status of all water bodies, including marine waters up to 1.85 kilometre off coast by 2015.
- The Marine Strategy Framework Directive (MSFD) (2008/56/EC), which aims to protect the EU marine environment in an effective way. A good environmental status of the EU marine waters should be achieved by 2020. Within this directive European Marine Regions will be established based on geographical and environmental criteria. To meet the requirements of this directive each EU member State is committed to develop strategies for their marine waters.

The decisions taken in the framework of these legal instruments should be based on sound scientific knowledge and it has been advocated that in the marine management, habitat suitability maps, biological valuation maps and biodiversity maps are suitable instruments

for communication with policy makers and marine managers (Derous *et al.*, 2007; Degraer *et al.*, 2008; Fraser *et al.*, 2008; Willems *et al.*, 2008).

A world to discover

Whereas land biodiversity patterns and the factors influencing this biodiversity are known for numerous taxa, our understanding of global biodiversity in the sea is more limited (Tittensor *et al.*, 2010). The ocean surface hides a diverse world which can only be discovered by sampling or diving at sea. This is an expensive and labour intensive task since adequate equipment and research vessels are needed. Therefore, it is important not only to collect but also to preserve valuable scientific data. In the 20th century, a lot of effort was done to discover the unknown marine world: sampling campaigns were organised, thousands of species were described, and information on species diversity and communities in the oceans was gathered. This information was often fragmented, but recently a lot of effort has been put into compiling this information in databases (Table 1.1). This is achieved by gathering researchers in regional or global consortia such as MarBEF and Census of Marine Life. In this way a better insight in the biodiversity of the oceans and the world as a whole can be achieved. This wealth of information gives new opportunities to explore and study the biodiversity of the oceans.

Project name	Database content	Number of species*	Website
CBOL iBOL	DNA barcode sequences of marine and terrestrial species	93 543 Target 2015: 5 000 000	http://www.barcodeoflife.org/ http://ibol.org/
Encyclopedia of Life	Marine and terrestrial species	Target: 1 900 000	http://www.eol.org/
Catalogue of Life	Marine and terrestrial species	1 333 403	http://www.catalogueoflife.org/col/
GBiF	Marine and terrestrial species	919 873	http://www.gbif.org/
Itis	Marine and terrestrial species	518 498	http://www.itis.gov/
WoRMS	Marine species	207 762	http://www.marinespecies.org/
OBIS (Census of Marine Life)	Marine species	114 879	http://www.iobis.org/
ERMS	Marine species	31 000	http://www.marbef.org/data
NeMys	Marine and terrestrial species	14 945	http://nemys.ugent.be
MarBOL	DNA barcode sequences of marine species	6 199	http://www.marinebarcoding.org/
MANUELA	Marine meiobenthic species	1 250	http://www.marbef.org/projects/Manuela/

*Table 1.1. Examples of biodiversity databases available on the web (*on February 1st 2011).*

Need for accurate models to understand and protect the marine habitat

The disclosure of this large amount of data helps in understanding the complexity of marine ecosystems, and modelling may serve both in the understanding and protection of this habitat. Mathematical models can reveal the factors influencing the biodiversity and may in this way contribute to the understanding of the structure of marine communities. On the other hand, there is a need to delineate areas which need to be preserved to protect the diversity and resilience of the seas. The designation of these protected areas should be based on sound scientific models. Delineating a protected area involves spatial multi-criteria-analysis (Villa *et al.*, 2002; Pomeroy *et al.*, 2004) encompassing a vast number of criteria such as geological features, diversity and composition of benthic and pelagic communities, potential human use of the area (tourism and fisheries), the protection of specific species, pollution status of the area, and many more (Villa *et al.*, 2002). Biodiversity maps and habitat suitability maps should thus be considered during the decision process. Given the importance of these models, it is crucial that these models are beyond discussion and all potential modelling pitfalls should be tested for, and avoided. In this way, these models can really contribute to understanding and preserving biodiversity. In this thesis, we investigate how the construction of erroneous or non-significant models caused by spatial autocorrelation and preferential sampling can be avoided.

BIOLOGICAL DATA

Here, we will focus on the marine benthos, with emphasis on free-living nematodes and the distribution of two macrobenthic species: an ecosystem engineer, the polychaete *Lanice conchilega* and an invasive species for the North-East Atlantic Area, the bivalve *Ensis directus*. The marine benthos encloses all those species associated with the sea floor. This diverse community can be divided in different groups according to size, location and type. Generally, benthic organisms are grouped according to size: macrobenthos (i.e. benthic organisms retained on a sieve with 1 mm mesh size), meiobenthos (i.e. metazoan organisms passing a 1 mm sieve but retained on a sieve of 38 µm) and microbenthos, (i.e. microscopic benthos passing a 38 µm sieve). The lower size of 38 µm for the meiobenthos may vary: sizes of 32 µm, 44 µm and 63 µm are also applied (Giere, 2009).

Nematodes

In this thesis we focus on a phylogenetic group within the meiobenthos: the Nematoda. The nematodes or roundworms are the most diverse phylum of pseudocoelomates, and one of the most diverse of all animals. Over 26 000 species have been described, of which over 16 000 are parasitic and more than 4 000 are free-living marine nematodes (Hugot *et al.*, 2001) (Table 1.2). It has been estimated that the total number of nematode species might be approximately 500 000 (Hammond, 1992) or even 1 000 000 (May, 1988) and they could be the second most diverse group after the Arthropoda (Hugot *et al.*, 2001). Nematodes are not

only highly diverse, but are often complex and biologically specialised metazoans (De Ley, 2006). Free-living nematodes represent a high diversity in many benthic environments in terms of species numbers (Heip *et al.*, 1985): more than 50 species are commonly found in a single 10 cm^2 core. In meiofaunal samples, nematodes are usually the dominant taxon both in abundance and in biomass (Giere, 2009).

Nematode communities are very useful as indicator organisms (Bongers and Ferris, 1999; Kennedy and Jacoby, 1999; Geetanjali *et al.*, 2002) in the assessment of sediment quality and pollution status of the environment (Schratzberger *et al.*, 2000a) for numerous reasons: they show a wide distribution from pristine to extremely polluted habitats, they do not rapidly migrate from stressful conditions; they respond rapidly to disturbance and enrichment; and they show a clear relationship between structure and function which can be deduced from the mouth cavity and the pharynx (Bongers and Ferris, 1999). In addition, owing to their interstitial life style, biogeochemical properties of the sediment have a strong influence on the diversity and the composition of nematode assemblages (Heip *et al.*, 1985; Steyaert *et al.*, 1999; Schratzberger *et al.*, 2000a; Vanaverbeke *et al.*, 2011).

Life style	# of species
Free-living marine	4 070
Free-living terrestrial	6 610
Plant parasites	4 110
Invertebrate parasites	3 500
Vertebrate parasites	8 360
Total	26 650

Table 1.2. Number of described nematodes (Hugot et al*., 2001).*

Moreover, biodiversity loss of nematode communities might be associated with exponential reduction of the ecosystem function (Danovaro *et al.*, 2008). Free-living nematodes fulfil many different functions in the sediment. Wieser (1953) divided nematodes in four trophic groups according to the shape of their buccal cavity. Selective and non-selective deposit feeders, epigrowth feeders and predators/omnivores were discerned. However, this classification is a simplification of the complex and diverse feeding patterns in nematodes. Some species are facultative predators or feed on ciliates (Moens and Vincx, 1997); other species show switches in feeding behaviour, depending on the available food (Giere, 2009) or the ontogenetic age (Lorenzen, 2000). Most nematodes are nowadays considered selective feeders (Giere, 2009). They can selectively differentiate between prey organisms (Moens *et al.*, 2000) and even between bacteria (Moens *et al.*, 1999). It has been postulated that the high diversity within a nematode community is caused by this food partitioning and resulting niche separation (Heip *et al.*, 1985; Moens and Vincx, 1997; Moens *et al.*, 1999).

These characteristics make the Nematoda a perfect phylum to study community patterns on a broad and local scale.

Macrobenthic species

In this thesis we also focus also on two macrobenthic species: *Lanice conchilega* and *Ensis directus.*

Lanice conchilega (sand mason) is a tube building polychaete and is a species with a wide spread bathymetrical (0-1900 m) and geographical range (Hartmann-Schröder, 1996). Conservation of the species as such is therefore not the main issue here. However, the habitat built by dense aggregations of the species is considered to be a reef (Rabaut *et al.*, 2009). The species changes its direct environment considerably (Rabaut, 2009) as it is considered to be an important ecosystem engineer (*sensu* Jones *et al.*, 1994). The worm builds linear tubes consisting of coarse sand grains cemented with mucus (Jones and Jago, 1993) which can reach a diameter of 5 mm and a length of 65 cm (Ziegelmeier, 1952). The tube is located mainly in the sediment, and only one to four centimetres protrude in the water column. This species has the ability to build dense aggregates and patches with more than 1500 ind.m^{-2} are not uncommon (Zühlke, 2001). These aggregations change the local sedimentary and hydrodynamic environment and the tubes themselves compact the sediment and increase the rigidity (Jones and Jago, 1993). This altered habitat induces changes in the benthic community, resulting in an increase of both macrobenthic abundance and diversity (Callaway, 2006; Rabaut *et al.*, 2007; Van Hoey *et al.*, 2008). Moreover, the *L. conchilega* reefs have a high functional value (Godet *et al.*, 2008) and are related to higher densities of juvenile flatfish such as *Pleuronectes platessa* (plaice) (Rabaut *et al.*, 2010). Recent research indicated the value of the species as a bio-irrigator, which pumps oxygen in the sediment. This mechanism can contribute to the mineralisation and denitrification process in the sediment (Braeckman *et al.*, 2010) and creates an extended habitat for nematodes (Braeckman *et al.*, 2011).

Thus, *L. conchilega* can be considered to be a valuable species in a conservation context (Van Hoey, 2006; Godet *et al.*, 2008; Rabaut *et al.*, 2009) and the development of habitat suitability models for this species has been advocated (Rabaut, 2009).

Ensis directus (Atlantic jackknife, American jackknife clam or razor clam) is a large edible bivalve species. It is indigenous to the Atlantic coast of Canada and North-America and prefers muddy, fine sand with small amounts of silt (Beukema and Dekker, 1995; Kennish *et al.*, 2004) and is found in the intertidal or subtidal zones (Mühlenhardt-Siegel *et al.*, 1983; Swennen *et al.*, 1985). It can burrow very quickly (Swennen *et al.*, 1985), and is also able to swim (Drew, 1907). It is probably introduced into Europe as larvae in ballast water of a ship crossing the Atlantic around 1978 (von Cosel, 1982). The first strong year class occurred in the German Bight in 1979 (von Cosel, 1982). Since then it has spread across the Dutch and Belgian coast. The first Belgian observations date from 1987 (Kerckhof and Dumoulin, 1987) and since its arrival, it has become the bivalve with the highest biomass in several areas along the coast (Tulp *et al.*, 2010). The species can occur in high densities (i.e. bivalve banks) and densities of 1000-2000 ind.m^{-2} are not uncommon (Armonies and Reise, 1999; Tulp *et al.*, 2010). These banks show a patchy distribution, but these patches are not permanent and

in Europe prominent events of mass mortality in late winter or early spring have been observed (Armonies and Reise, 1999). The potential distribution of this invasive species and its potential harmful effect on the natural community in Europe is largely unknown. This invasive species may compete for space and food with indigenous species and dense populations may change the community structure of the benthic fauna (Gollasch *et al.*, 1999). Therefore, habitat suitability modelling can be useful for creating insight in the ecology and the possible distribution of the species. Also, fisheries have shown interest in fishing this bivalve and there might be a link between sea ducks (i.e. Common Scoter (*Melanitta nigra*)) and high densities of *E. directus* (Houziaux *et al.*, 2010). Therefore, distribution maps reflecting densities will be created by applying geostatistics.

STUDY AREA

The study area changes throughout the thesis since different databases and different subsets of these databases have been used. However, the main areas of interest are the continental shelf area of the Southern Bight of the North Sea (referred to as SBNS) (Fig. 1.2) and the Belgian Part of the North Sea (BPNS)[1].

The North Sea is a shallow sea located between Great Britain, Scandinavia, Denmark, Germany, the Netherlands and Belgium. The SBNS is delimited to the North by the thermal stratification of the water column during summer (around 54°N) and to the South by the Strait of Dover between Great Britain and France. The area has a maximum depth of about 54 m (Fig. 1.2) and is characterised by strong semi-diurnal tidal currents (up to about 1 $m.s^{-1}$) and frequent strong winds. This results in a well mixed water column throughout the year (Lee, 1980). The area is characterised by a complex system of sand banks which follow the residual current and are therefore oriented parallel to the coast (Muylaert *et al.*, 2006). The net bottom shear stress is directed to the North (Pingree and Griffiths, 1979).

The seabed sediments consist mainly of fine to medium sands (125-500 µm) (Verfaillie *et al.*, 2006). The Eastern part of the Belgian coast is characterised by high concentrations of silt-clay (Fig. 1.2). The origin and the formation of these silt-clay deposits in front of the coast are explained by the neap-spring tidal cycles and the different sources of Suspended Particulate Matter (SPM). These sources are mainly the erosion of exposed clay layers and the import of SPM from the Strait of Dover (Fettweis and Van den Eynde, 2003).

The SBNS receives carbon and nutrients from river inputs (mainly from the Rhine, Meuse and Scheldt), atmospheric deposition, and exchanges with the Atlantic Ocean through the English Channel (Baeyens *et al.*, 2007). The sources of nutrients and carbon are mostly linked to anthropogenic activities (agriculture, industries, domestic wastewater) (Baeyens *et al.*, 2007). These nutrients give rise to pronounced phytoplankton blooms in spring and late summer. The spring bloom is dominated by two major phytoplankton groups, diatoms and *Phaeocystis* (van der Zee and Chou, 2005). Chl *a* deposition on the sea floor is directly

[1] In Chapter 2 the BPNS is referred to as the Belgian Continental Shelf (BCS)

related to these phytoplankton blooms (van Oevelen *et al.*, 2009). The organic matter in the water column can enter the sediment through physically mediated input (advective injection and passive deposition), or by benthic organisms which actively filter and deposit organic matter from the water column on and into surface sediments (Kautsky and Evans, 1987; Kotta *et al.*, 2005). Mineralisation of this newly settled organic matter can induce hypoxic or even anoxic conditions in the sediment (Graf, 1992).

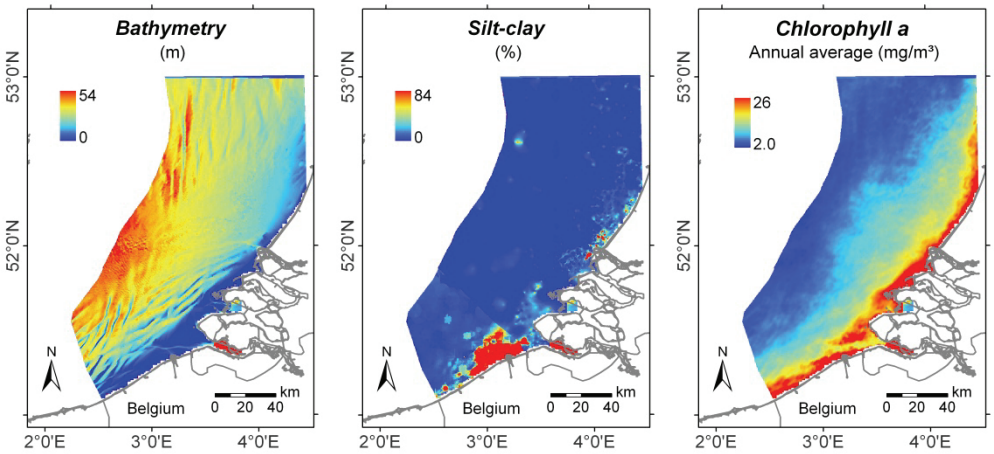

Fig. 1.2. Bathymetry, silt-clay fraction of the sediment, and average chlorophyll a *in the water column of the Southern Bight of the North Sea*[2]

BIODIVERSITY

Biodiversity indices

Biodiversity is defined by Glowka *et al.* (1994) as '*The variability among living organisms from all sources including, among other things, terrestrial, marine, and other aquatic ecosystems and the ecological complexes of which they are a part; this includes diversity within species, between species and of ecosystems.*' Biodiversity can be expressed in many different ways: it can be linked to species richness, evenness, taxonomical diversity, genetic diversity, functional diversity or other features of the species community. The most common expression of biodiversity is the number of different species in a given area (species richness). However, this estimate of biodiversity is strongly influenced by sample size, and a number of statistical techniques have been developed to correct for this: estimators for total species richness (Chao, 1984; Chao, 1987) and evenness (Chao and Shen, 2003), the expected species richness (Sanders, 1968; Hurlbert, 1971; Simberloff, 1972), and taxonomic diversity indices (Clarke and Warwick, 1998). Depending on the scale α-, β- and γ-diversity can be distinguished: α-diversity is the biodiversity within a particular area (e.g. within the

[2] source: Renard Centre of Marine Geology (RCMG, www.rcmg.ugent.be) of Ghent University and the Hydrographic Service of the Royal Netherlands Navy and the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management for the oceanographic and sedimentological data.

area of a core sample). β-diversity is a measure of biodiversity which compares the species diversity between ecosystems. This involves comparing the number of taxa that are unique to each of the ecosystems and gives a view on species turnover across habitats. γ-diversity refers to the total biodiversity over a large area or region (Whittaker, 1972). The 'intrinsic' diversity of a community is given by its α-diversity. Thus, an area with higher α-diversity may be considered more important than one with lower α-diversity values, for conservation purposes (Hernández-Stefanoni and Ponce-Hernandez, 2004). But the contribution of an area to the overall γ-diversity is also depending on the β-diversity. Therefore, areas with lower α-diversity may still be important to conservation management because of their contribution to the total diversity of the area. Developing an additional map with an estimation of the β-diversity would be interesting from a conservational point of view (Samson and Knopf, 1982). However, available approaches to predict β-diversity are hampered by the twofold scale dependence of β-diversity, owing to the size of sampled units as well as their mutual distances (Feilhauer and Schmidtlein, 2009). Moreover, for the SBNS more species have a sample specific name (e.g. *Araeolaimoides* sp.1 MV, with MV the reference to the data supplier) than there are species with an accepted name. On average about 17% of the species in a sample are sample specific. Thus, these species can be used to estimate the α-diversity of a sample, but they cannot be used for estimating β-diversity. Moreover, the maps of the environmental variables should be transformed in gradient maps which will further increase the error on the model. For these reasons an error proliferation on the prediction of β-diversity might be expected. Therefore, to keep the error rate as low as possible, we focussed on α-diversity of the core samples.

Biodiversity patterns

Local or regional variation in biodiversity has fascinated many researchers. Which processes and factors may explain those differences in biodiversity? Darwin described evolutionary processes and natural selection as the driver behind species differentiation (Darwin, 1876). The distribution of these species is however not homogeneous and the spatial and temporal variation in benthic biodiversity reflects not only evolutionary processes but also ecological processes which operate at different spatial and temporal scales (Levin *et al.*, 2001). Many hypotheses have been postulated to offer an explanation for these spatial variations in biodiversity (Table 1.3). There is a considerable overlap between these theories, and some operate on a large temporal and spatial scale, some on a small scale. These small-scale processes are hierarchically embedded in processes taking place on a larger scale. The local processes include competition, exclusion, facilitation, resource partitioning, disturbance of the physical environment and physiological tolerance (Etter and Mullineaux, 2001). On a larger regional scale, environmental factors are important for structuring benthic communities (Fig. 1.3).

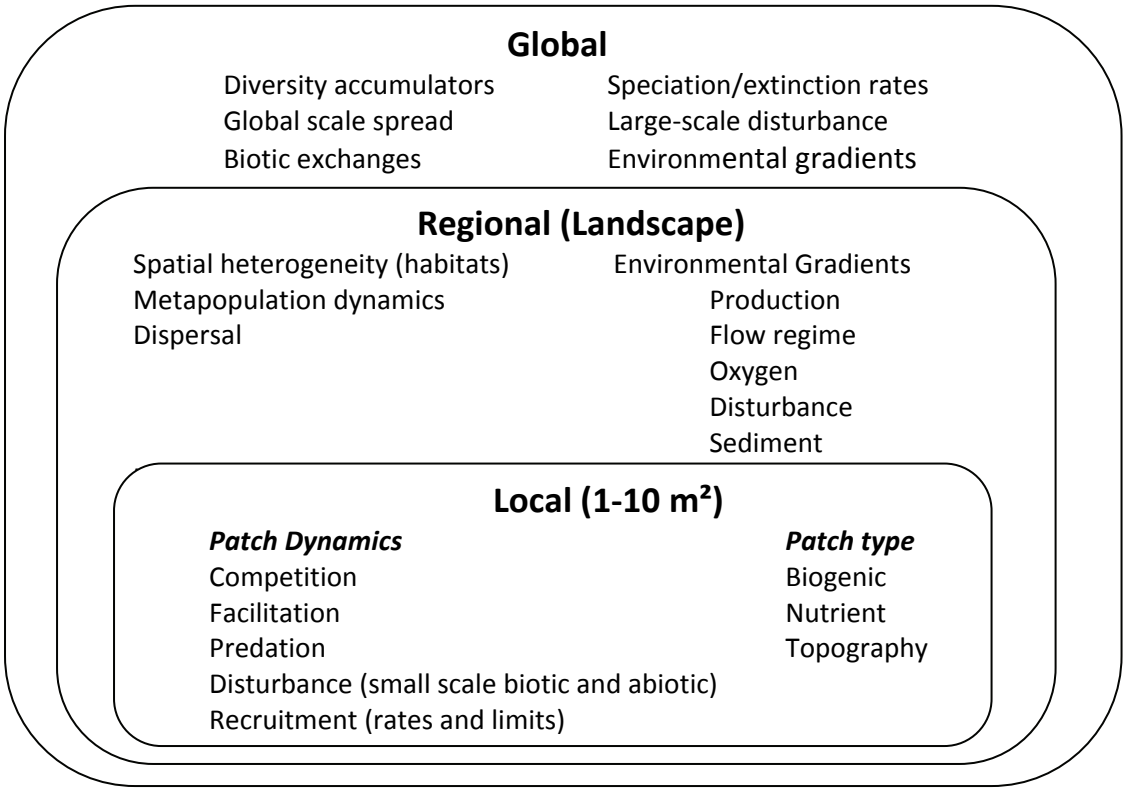

*Fig. 1.3. Processes regulating species diversity at local, regional and global scales (*from *Levin* et al.*, 2001)*

their response to environmental factors (Ricklefs and Shluter, 1993). On a large scale, metacommunities are formed: metacommunities are local communities connected by the dispersal of multiple species (Hubbell, 2001). Leibold *et al.* (2004) suggested four types of metacommunities: the species sorting, source-sink dynamics, the neutral model and patch dynamics type. Here, research on metacommunities is impaired, since little is known about the dispersal capacities of nematode species. Nevertheless, we touch upon two aspects in the discussion: patch dynamics and species sorting.

In this study, the importance of both local and regional factors on biodiversity will be addressed: local processes involve species interactions and will be investigated by null models involving 'species assembly rules'. On a larger scale, environmental factors are assumed to be important. The influence of these abiotic factors on biodiversity and species distributions will be investigated with techniques such as artificial neural networks, generalised least squares and habitat suitability modelling.

Species assembly rules

Species are often unevenly distributed in space and form distinct species communities. One of the fundamental questions in ecology is how unexpectedly strong species associations are created in nature. Evolutionary ecologists are concerned about the question: how do speciation and extinction form a species pool, while community ecologists focus rather on

the question: how are communities formed from this species pool (Weiher and Keddy, 2001). The latter question is addressed by the 'assembly rules'.

Hypothesis	Description
Competition (Dobzhansky, 1950; Dayton, 1971; Grassle and Sanders, 1973; Diamond, 1975)	Species evolve through competition, occupy their own niche and develop therefore specific morphology.
Biological disturbance hypothesis (Paine, 1969; Dayton and Hessler 1972)	Predator-controlled diversity reduces abundance of competitors, maintains resource availability and prevents competitive exclusion.
Intermediate disturbance hypothesis (Connell, 1978) and the dynamic equilibrium (Huston, 1979)	Diversity is maximised at an intermediate frequency and/or magnitude of disturbance
Habitat heterogeneity hypothesis (MacArthur and MacArthur, 1961; Bazzaz, 1975)	Spatial heterogeneity leads to faunal diversification. Specialisation arises through microhabitat exploitation and niche partitioning.
Productivity-diversity hypothesis (Connell and Orias, 1964; Yount, 1956; Rex, 1981)	Positive (Connell and Orias, 1964), negative (Yount, 1956) and optimum relations (Rex, 1981) between productivity and diversity have been found.
Island biogeography (MacArthur and Wilson, 1967)	Isolated islands foster the evolution of new species.
Stability-time hypothesis (Connell, 1978)	Stressed and young environments will have less species than old and/or stable environments.
Theory of climatic stability (Klopfer, 1959)	Regions with stable climates allow the evolution of finer specialisations.
Historical explanation (Jackson, 1992)	Past geological events (e.g. sea level rise) may explain present diversity patterns.
Latitudinal diversity hypotheses (Fischer, 1960; Rohde, 1992; Rex *et al.*, 1993)	Diversity decreases with increasing latitude. Several hypotheses have been postulated to explain this pattern.

Table 1.3. Selection of hypotheses and influences explaining spatial differences in biodiversity.

There are mainly two confronting ideas explaining the assembly of communities: the island paradigm and the trait-environment paradigm (Weiher and Keddy, 2001). The island paradigm relates the structure of communities to dispersal, competition, immigration and extinction, while the trait-environment paradigm considers the environmental factors as the main driving force in structuring communities. As mentioned before, the influence of the environment on diversity and the occurrence of species are in this thesis investigated on a larger geographical scale. Therefore, in this thesis, the term 'species assembly rules' will relate to the 'island paradigm' and thus species interactions. This idea was first introduced by Diamond (1975). He argued that interspecific competition between species occupying similar niches results in a non-random pattern of species distributions. More specifically, some species pairs may never be found together due to competitive exclusion, forming a perfect checkerboard pair. His rules have been strongly disputed and there has been a proliferation of studies promoting, refuting and testing these ideas (Connor and Simberloff

1979; Diamond and Gilpin, 1982; Weiher and Keddy, 1995; Bell, 2000; Weiher and Keddy, 2001; Hubbell, 2001; Bell *et al.*, 2006; Purves and Turnbull, 2010). It has been shown that neutral factors such as birth, death, random dispersal and the total number of organisms in the community may even result in these non-random patterns (Bell, 2000; Hubbell, 2001). However, recent work shows that neutral processes alone cannot explain the observed community patterns (Bell *et al.*, 2006; Purves and Turnbull, 2010).

Most of the community assembly research concentrated on terrestrial studies (Gotelli and McCabe, 2002; Ribichich, 2005) or marine macrobenthos (Pagliosa, 2005). In contrast, Nematoda have received considerably less attention. The main advantages of using nematodes lie in their resistance to disturbance and the vast number of species found in small volumes of sediment.

While environmental variables structure nematode communities on a large scale (Vanaverbeke *et al.*, 2011), species interactions may become important on a smaller scale (Joint *et al.*, 1982; Steyaert *et al.*, 2003). Therefore, we focussed our search for assembly rules on repeated samples at the same location at the same moment in time (further referred to as 'replicate samples'). Replicate samples are generally collected within a small area where differences in environmental variables are small relatively to the environmental differences on a larger scale. Thus, the presence of non-random communities is tested on replicate samples with the use of null models.

It is important to note that the goal of these theoretical models is only to recognise non-random community patterns (Gotelli and Graves, 1996). Revealing the cause of these non-random patterns can only be established by experimental set-ups (Gotelli and Graves, 1996; Gotelli and McCabe, 2002).

Biodiversity and the environment

Levin *et al.* (2001) described six main factors structuring biodiversity in marine benthic environments: boundary constraints, sediment heterogeneity, productivity and food supply, bottom-water oxygen concentrations, sea currents and catastrophic disturbance. Their overview focused mainly on deep-sea macrobenthic species, but studies on the nematode communities of shallow seas revealed that many of these factors are important as well for structuring nematode communities (Heip *et al.*, 1985; Vanreusel, 1990; Vincx *et al.*, 1990; Steyaert *et al.*, 1999; Vanaverbeke *et al.*, 2002; Vanaverbeke *et al.*, 2011).

As mentioned before, biodiversity can be expressed in different ways. In this study we focused on diversity indices related to species richness, evenness and taxonomic richness.

Choice of modelling technique

The relationship between the diversity aspects and the environment may be complex and difficult to model and commonly used mathematical models may be inadequate (Lek and Guégan, 1999). In modelling different approaches can be used depending on the aim of the

model. Multiple linear regression (MLR) is one the most frequently used predictive methods in ecology. The popularity of MLR and linear models in general lies in its ease of use and its capacity to give explanatory results, as the coefficients of the environmental variables provide straight-forward information about their relative importance and they can give some measures of confidence about the estimated coefficients. However, linear models are based on the *a priori* selection of suitable functions and algorithms, such as linear relations in traditional linear modelling or a link linear relation in generalised linear models (Park and Lek, 2005). Therefore, linear models cannot deal with non-linear relationships between dependent and independent variables, unless complex data transformations are performed (Gevrey *et al.*, 2003). Gevrey *et al.* (2003) stressed another drawback of linear regression techniques: only variables with statistically significant coefficients are analysed, for which it may lack resolution. That is why the use of more complex supervised learning techniques is justified in ecology where the relationships between variables are principally non-linear (Lek and Guégan, 1999; Gevrey *et al.*, 2003). In data mining, a large range of modelling techniques are available: support vector machines, classification and regression trees, random forests, artificial neural networks (ANNs) and many more (Izenman, 2008). ANNs are known as powerful computational tools. They can be used for many purposes: classification, pattern recognition and modelling based on empirical data. Most of these tasks can also be performed by conventional statistics. However, ANNs often provide a more effective way to deal with problems that are difficult, if not intractable, for traditional computation (Park and Lek, 2005). As mentioned before, linear models are based on the *a priori* selection of suitable functions and algorithms. For ANNs no such *a priori* selection is needed. ANNs with a single layer can approximate any mathematical function and they can treat complicated problems, even if the data are imprecise and noisy. Their superior modelling capacity is also apparent from the 35 papers we analysed: in 34 out of 35 papers ANNs outperformed conventional linear models. Moreover, their implementation is not precluded by the theoretical distribution shape of the data (Bishop, 1995). If enough data is available and the architecture is properly selected, ANNs provide optimal solutions for any relation between the dependent and the independent variables. In this way unknown or unsuspected relations can be revealed. The relation between the sediment characteristics and the diversity of nematode communities has been widely reported, however the relationship on a large scale was not yet established. Moreover, our goal was to find out which aspects of biodiversity are best explained by the large scale environmental variables. To allow a fair comparison between the different diversity indices, a powerful and flexible tool, which is not subject to modelling constraints, was needed. An additional advantage of using ANNs is that the optimisation process to find an optimal architecture can be automated. This is especially appropriate when several dependent variables need to be modelled. However, a drawback of the flexibility of the ANNs is that the models are prone to overfitting. Therefore, special attention was paid to find the optimal network which does not overfit. Moreover, ANNs are often seen as 'black boxes' which means that the contribution of each input variable to the model output is hidden in the model layers and is difficult to disentangle from the network

(Lek *et al.*, 1996a). However, to reveal the contribution of each environmental variable to the resulting models, we applied three methods: two known in literature: the Perturb method (Yao *et al.*, 1998; Gevrey *et al.*, 2003) and the Profile method (Lek *et al.*, 1995, 1996a, b; Gevrey *et al.*, 2003). A third technique was developed to check the validity of the previous two.

MAPPING

Mapping biodiversity of regions

Biodiversity may vary considerably across and within regions. At this scale, environmental influences contribute to these biodiversity differences. The relation found between the environmental variables and the biodiversity can be used to create diversity maps by applying geostatistics. Geostatistics is a branch of statistics that allows the estimation of the values of a variable of interest at non-sampled locations. Although this approach is related to interpolation methods, it extends far beyond simple interpolation problems (Van Meirvenne, 2007). It consists of a collection of mathematical techniques dealing with the characterisation of spatial phenomena. The technique was originally developed to predict probability distributions of ore grades for mining operations (Matheron, 1963). Currently, geostatistics are applied in diverse disciplines including petroleum geology, oceanography, geography, forestry, environmental control and landscape ecology. More recently, it is also applied in the marine environment to map distribution patterns of marine species (Mello and Rose, 2005; Rios-Lara *et al.*, 2007) and biodiversity patterns (Reese and Brodeur, 2006).

Particularly interesting are the hybrid interpolation techniques. They rely on two techniques (a) interpolation relying on point observations of the spatial variable; and (b) interpolation based on regression of the target variable on spatially exhaustive environmental variables (maps) (Hengl *et al.*, 2007). One of these techniques is known as regression kriging (Hengl *et al.*, 2004). First a regression of the dependent variable with the environmental variables is applied and then it uses simple kriging to interpolate the residuals from the regression model. This allows the use of any regression method to correlate the dependent variable with the environmental variables (Hengl *et al.*, 2007).

Mapping a species' potential distribution

The diversity of a taxon across a region gives an estimation of the species-rich areas. However, this does not supply any information on the community composition. The presence of a single species may be important and reveal underlying structures. Especially keystone, indicator or umbrella species may hold more information about the community. Other species such as *Lanice conchilega* contribute to the diversity of the area by modifying the habitat as a habitat engineer (Callaway, 2006; Rabaut *et al.*, 2007; Van Hoey *et al.*, 2008) and forming reefs, while the distribution of invasive species such as *Ensis directus* may be of

interest to nature conservationists and fisheries. In that case habitat suitability models (HSMs) reveal information on where the species can potentially be found.

Choice of modelling technique

Different modelling techniques have been developed to estimate the potential habitat of a species. Depending on the type of input data, two types of HSMs can be discerned: those based on presence/absence data and HSMs based on presence-only data. Presence/absence data are commonly used for HSMs, but this includes some presumptions about the information. Often the absence of a species is not 100% sure for different reasons 1) not all organisms in a sample are identified, 2) species show patchy or ephemeral distributions and may not be present at the time of sampling although the habitat is suitable for the species, or 3) the species may not have obtained its full range because of a disturbed environment or because it is an invasive species. In such cases presence-only modelling techniques are preferred. Several presence-only modelling techniques have been developed: Bioclim (Nix, 1986), Domain (Carpenter *et al.*, 1993), GARP (Stockwell and Peters, 1999), Maxent (Phillips *et al.*, 2004) and ENFA or Biomapper (Hirzel *et al.*, 2002). Maxent has proven its better predictive capacities compared to other presence-only modelling techniques in several independent cases (Hernandez *et al.*, 2006, 2008; Hijmans and Graham, 2006; Pearson *et al.*, 2007; Sergio *et al.*, 2007; Carnaval and Moritz, 2008; Ortega-Huerta and Peterson, 2008; Benito *et al.*, 2009; Roura-Pascual *et al.*, 2009) and may compete with or even outcompete presence/absence modelling techniques such as boosted regression trees (BRT), generalised additive models (GAM), generalised linear models (GLM) and multivariate additive regression splines (MARS) (Elith *et al.*, 2006; Wisz *et al.*, 2008). These good predictive capacities have been attributed to the ℓ_1-regularisation which prevents the algorithm from overfitting. Other models often do not apply any form of regularisation, and this can cause the observed difference in predictive performance (Gastón and García-Viñas, 2011). Moreover, Maxent is a generative approach, rather than discriminative. This can be an inherent advantage when the amount of training data is limited (Phillips *et al.*, 2006). Research pointed out that the technique can be applied with as little as 5 sampling points (Pearson *et al.*, 2007). For the nematode species in the database occurrence data is often scarce. A software related advantage is that it allows computerising the calculation of thousands of HSMs by running batch-files. Therefore, Maxent was applied in this thesis to create habitat suitability models. In spite of these promising features, Maxent models seem to have two major drawbacks: the models may fail to make general predictions (Peterson *et al.*, 2007) and the models may be inaccurate in the presence of biased data (preferential sampling) (Phillips *et al.*, 2009). In this research, these drawbacks are tackled in various ways.

Mapping a common species

Species with specific habitat requirements and a small spatial range are generally easier to model than common species (Segurado and Araújo, 2004; Evangelista *et al.*, 2008) because widely distributed species are not restricted to specific habitats. However, knowing where to find high densities of a common species may be of interest to fisheries (e.g. *Ensis directus*) or for conservation purposes (e.g. *Lanice conchilega* reefs). Application of regression kriging techniques could be an option, but at least 100, and preferably 144 observations are required for reliable kriging (Webster and Oliver, 2007). Clearly, this represents some limitation to the applicability of regression kriging. Therefore, an alternative was investigated: instead of using presence-only or presence/absence data, a threshold was applied to the density or relative density of the species. The applicability of this methodology was tested for 6 common nematode species and applied to *Lanice conchilega*.

MODEL OPTIMISATION

Creating a model is generally speaking not difficult. There are ample examples of software modules which create a model within seconds if data is supplied in the correct format. However, creating a significant model which provides correct insight in the ecology of species and communities needs careful consideration. There are many pitfalls when it comes to creating a model (Ülgen *et al.*, 1996; Rosemann, 2006a; Rosemann, 2006b): pitfalls concerning the data, the modelling technique and model interpretation. Counteracting these pitfalls will result in more accurate and reliable predictive models.

Pitfalls concerning the data

Ecological and environmental data are spatial data, and spatial data may be subject to two potential problems: spatial autocorrelation and preferential sampling.

A characteristic of most spatial data is that it shows spatial autocorrelation (SA). In other words, if two samples are taken closely together they are more likely to resemble each other than if the samples were taken at a larger distance from each other (positive SA). Consequently, both samples are not independent and the statistical assumption of independence is violated. SA may amplify or blur true ecological relations and incorporating spatial autocorrelation may even invert observed patterns (Kühn, 2007). Moreover, SA may inflate validation statistics since the test localities are not spatially independent from the training data (Hampe, 2004) (see Paragraph below for more information on training and test set). The presence of spatial autocorrelation is established by calculating Moran's I (Moran, 1950). Accounting for SA is often done by spatially separating training and test set (Pearson *et al.*, 2007; Murray-Smith *et al.*, 2009).

When spatial data is collected, preferential sampling may occur, i.e. some areas are more frequently visited than others. A model is more likely to deviate from a random model when such collection bias occurs (Raes and ter Steege, 2007). If the sampling points correlate with

some environmental characteristics, these environmental characteristics may erroneously 'explain' the patterns found in the data, while their contribution to the observed pattern may be negligible. Ignoring preferential sampling can thus negatively affect model parameter estimation and prediction. Here, we analyzed the presence of preferential sampling by either plotting the declustered mean as a function of cell size (Van Meirvenne, 2007) or by the use of null models (Raes and ter Steege, 2007).

Pitfalls concerning modelling

Overfitting occurs when the overly complex model describes the random error or noise instead of the underlying relationship. In that case, the model will fit very well the training data but it will perform poorly when applied to new, unseen data. The model is therefore not able to generalise (Izenman, 2008). Reducing model complexity will often result in better generalisation capability. However, if the model is overly simple it will not be able to explain the variation in the data. Consequently model complexity often has an optimum. This optimum can be found by applying cross-validation or by using an independent validation set which is solely used at the completion of the modelling exercise. In case of ANNs, three sets are created: a training set, a validation set and a test set. During the refining of the model weights, the error on the validation set is monitored. Once the error of the validation set increases, the iteration process is stopped. This is an internal validation, thus the validation set is not an independent dataset. Therefore, a third set, the test set, is created. k-fold cross-validation is a technique to assess if a model is capable to generalise to unseen data. The data is split in k partitions, each partition is once used as a validation set, while the other (k-1) partitions are used to train the model (training data) (Olson and Delen, 2008). Choosing the number of folds is always a trade-off between more samples for testing or more samples for training the model. Generally a 10-fold cross-validation gives good results concerning the bias and variance on the accuracy estimation of the models (Kohavi, 1995). For Chapter 3 a 10-fold cross-validation could be applied since a lot of data were available (209 samples) and the fast Levenberg-Marquardt training algorithm is used (Beale *et al.*, 2010). However, in case of data and time limitations a three-fold or a five-fold cross-validation may be more appropriate (Goethals, 2005). For the habitat suitability modelling (Chapter 4 to 7) the number of data points varied between 5 and 106 and the modelling speed seriously dropped with increasing number of samples. Therefore, a lower number of folds were applied, both due to data and time limitations.

During the modelling, it is tempting to optimise and keep on optimising modelling parameters or look for further refinements. Increasing the complexity of the model, leads easily to being 'lost in best practice' (Ülgen *et al.*, 1996; Rosemann, 2006b). Moreover, if an immense amount of data is available, there is an almost unlimited choice in modelling possibilities. Setting clear targets and limitations should also be part of the modelling process (Ülgen *et al.*, 1996).

Pitfalls concerning the model output and interpretation

Statistical modelling involves finding a relation between the independent variables and a dependent variable. However, it is important to realize that correlation does not necessarily imply causality. In other words, if a mathematical relation is found between the dependent and the independent variable, this does not necessarily mean that the independent variable causes the changes in the dependent variable.

A model which can explain the variation in the data is only meaningful if the contribution of the environmental variables to the model output is known. More specifically, ANNs are generally known as 'black boxes'. Several techniques have been developed to reveal the contribution of each variable (Gevrey *et al.*, 2003). In this way the model can contribute to the ecological knowledge of the species. Having too much faith in the model output without knowing the limitations of the model is hazardous and may lead to false conclusions (Ülgen *et al.*, 1996). Coupling back to the ecological relevancy of the conclusions is a first step in counteracting this pitfall.

Explaining these limitations to potential end users should be part of the modelling process (Ülgen *et al.*, 1996).

DESCRIPTION OF THE DATA

Species data

The nematode data were retrieved from two databases: a database compiled within the UGent and the MANUELA database, which partly comprises the UGent database.

The UGent database was compiled within the framework of this thesis. In order to build a consistent database different quality controls were performed (Addendum 2). The database consists of historical data collected at Ghent University. It consists of 22 subsets with data collected in the framework of PhD research, BSc dissertations and funded research projects (Addendum 2). The complete database contains information on locations from all over the world, including the poles and tropical regions. However, the bulk of the data is collected at the SBNS. The database contains more than 206 000 species identifications and assembles data within the time frame 1971-2004. This database has the advantage that sampling and identification techniques are more similar for the data used in the analyses (see Addendum 2). This database is used in Chapters 3 to 6.

The MANUELA database was compiled within the EU Network of Excellence MarBEF. MANUELA was a Research Project focusing on the meiobenthic assemblages. The MANUELA database captures data on meiobenthos on a broad European scale (Vandepitte *et al.*, 2009). Table 1.4 gives an overview of the main characteristics of the UGent and the MANUELA database. The UGent database is included in the MANUELA database.

The macrobenthos data was retrieved from the MacroDat database and completed with new data in the case of *Ensis directus*. The MacroDat database is compiled at Ghent

University, with data of over 1275 stations and more than 640 000 species identifications collected between 1971 and 2008 (Degraer *et al.*, 2003a).

	# of subsets	# of stations	# of nematodes identified	# of nematode species	Time frame
UGent database	22	279	206 000	1975 (taxon level) 588 (species level)	1971-2004
MANUELA-database	82	1213	> 1 500 000*	1484 (taxon level) 951 (species level)	1966-2006

*Approximated value since counts were not always available

Table 1.4: Main characteristics of the UGent and the MANUELA database.

Environmental data

The environmental data was, if possible, retrieved from the databases. This is in fact the most accurate data, since it links the biotic data directly with the environmental data. However, in many cases this data is unavailable or incomplete and should be replaced by data from environmental maps. The advantage of maps is that a value can be drawn for each pixel within a given area that is mapped. A drawback is that these data are less accurate compared to data collected in the field. The maps have a resolution of 199 m x 199 m for the BPNS and 249 m x 249 m for the SBNS. Table 1.5 gives an overview of the available maps for both areas.

The available maps were acquired in two ways: by remote sensing or by interpolating field data and modelling. The data acquired by remote sensing covers data on total suspended matter and chlorophyll *a* in the water column (Park *et al.*, 2006). The data was collected by the MERIS spectrometer on board of the Envisat satellite within the Belcolour project. Eighty chlorophyll *a* maps and 90 total suspended matter maps were gathered in the time frame 2003-2005. These maps were reduced to three biologically relevant maps: the minimum, maximum and average values.

The second group of maps was derived from point sampling at sea or from modelling. The water current properties were modelled by the Management Unit of the North Sea Mathematical Models and the Scheldt estuary (MUMM). The sediment characteristics, such as median grain size and the silt-clay fraction were supplied by the Renard Centre of Marine Geology (RCMG) at Ghent University (Verfaillie *et al.*, 2006) and TNO Built Environment and Geosciences-Geological Survey of the Netherlands (TNO). The bathymetrical data were provided by the Ministry of the Flemish Community Department of Environment and Infrastructure, Waterways and Marine Affairs Administration and completed with data from the Hydrographic Service of the Royal Netherlands Navy and by the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management (RNLN). The only map representing anthropogenic effects on the sea

floor reflects information on the intensity of sand extraction. This data was supplied by the Federale Overheidsdienst Economie (FOE).

Variable type	variable	BPNS	SBNS	Source
Anthropogenic	Intensity of sand extraction	x		FOE
Biochemical	Average total suspended matter	x	x	Belcolour
	Maximum total suspended matter	x	x	Belcolour
	Minimum total suspended matter	x	x	Belcolour
	Average chlorophyll content	x	x	Belcolour
	Maximum chlorophyll content	x	x	Belcolour
	Minimum chlorophyll content	x	x	Belcolour
	Average salinity	x		Belcolour
	Maximum salinity	x		Belcolour
	Minimum salinity	x		Belcolour
Current prop.	Minimum bottom shear stress	x		MUMM
	Mean bottom shear stress	x		MUMM
	Maximum bottom shear stress	x		MUMM
	Size of the residual currents	x		MUMM
	Maximum depth-averaged current velocity	x		MUMM
	Magnitude of the residual transports	x		MUMM
	Residual currents	x		MUMM
	Residual transports	x		MUMM
	Tidal amplitude	x		MUMM
	Maximum current velocity at the bottom layer	x		MUMM
	Average current velocity at the bottom layer	x		MUMM
Oceanographic	Water depth	x	x	RCMG & RNLN
	Slope of the sea bottom	x		RCMG
	Bathymetric Position Index (1600 m range)	x		RCMG
	Bathymetric Position Index (240 m range)	x		RCMG
	Rugosity of the bottom	x		RCMG
	Orientation of the slope of the bottom	x		RCMG
Sediment	Median grain size	x	x	RCMG & TNO
	Gravel content	x		RCMG
	Sand content (63 µm - 2 mm)	x		RCMG
	Silt-clay content (0-63 µm)	x	x	RCMG & TNO

Table 1.5. Environmental variables available for the BPNS and the SBNS.

AIM AND OUTLINE OF THE THESIS

The overall aim of this research is to develop powerful statistical models which cope with modelling pitfalls which are typical for spatial data and data assembled from various sources, but are often ignored (Dormann, 2007). In this respect issues such as spatial autocorrelation, sampling bias (preferential sampling) and sampling effort are addressed. The aim of this thesis is two-fold: 1) adapting the currently used modelling process in such a way that potential pitfalls for a given dataset are revealed and circumvented; 2) getting insight in the ecology and biodiversity of the taxa under study on a small and large spatial scale. More specifically the diversity of nematode communities and the factors contributing to this diversity are modelled. Besides, diversity and community modelling this research focuses on habitat suitability modelling of nematode species and two macrobenthic species.

In this thesis, modelling focuses on statistical modelling. Therefore, different statistical modelling techniques such as artificial neural networks, geostatistics and maximum entropy modelling were used for the reasons mentioned before. These modelling techniques were combined with techniques such as cross-validation and null models to improve model quality and to address the previously mentioned pitfalls in order to create significant models which are able to generalise to unseen data (Segurado *et al.*, 2006; Dormann, 2007; Raes and ter Steege, 2007).

In Chapter 2 the potential role of species assembly rules on a small scale are investigated by applying null models to the original dataset. More specifically, we addressed the question whether the species composition in replicate samples is different from a random species distribution based on the local species pool and what might cause any possible non-random patterns. A routine was developed in Matlab to evaluate if the species composition between replicate samples is significantly different from a random distribution from the local species pool.

Chapter 3 explores the relation between biodiversity indices and environmental variables. Artificial neural networks are often considered to be 'black boxes', not revealing insight in the relation between dependent (i.e. biodiversity indices) and independent (i.e. environmental variables) variables. Therefore, two existing (Perturb and Profile) and one new technique (Modified Profile) were applied to reveal these relationships. As such, we were able to investigate the factors related to the observed diversity patterns.

In Chapter 4 different modelling and kriging techniques are compared to find the best map of nematode diversity for the SBNS. The use of replicate samples in geostatistical modelling is explored. These replicate samples may give a good insight in the local variation of the biodiversity caused by sampling errors and variability at small distances and this information may contribute to the quality of the final model.

Chapter 5 explores the influence of preferential sampling, spatial autocorrelation and overfitting on habitat suitability models. Null models were applied before to reveal the presence of preferential sampling (Raes and ter Steege, 2007), but they are equally useful in

revealing spatial autocorrelation and overfitting. As a result habitat suitability models for more than 100 nematode species were developed. These resulting models can be found in Addendum 3.

Well performing habitat suitability models are generally difficult to develop for common or generalist species with no specific habitat requirements. However, in some cases it is relevant to know where these species can be found in high densities. Therefore, we investigated the influence of the relative abundances on the outcome of habitat suitability models of six nematode species in Chapter 6. Do the high relative abundances relate to specific habitats?

Chapter 7 illustrates two applications of these modelling techniques to *Lanice conchilega* and *Ensis directus*.

The final chapter, Chapter 8, holds general conclusions and possible opportunities for future research. A theoretical introduction to artificial neural networks, geostatistics and maximum entropy modelling is supplied in Addendum 1. The codes developed in Matlab and R can be found in Addendum 5.

REMARK

Apart from the introduction, Chapter 7 and the general discussion, this thesis is a compilation of research articles which have been published or will be submitted to peer reviewed journals. This may result in overlapping information about the data and modelling techniques in the different chapters, but it also means that the chapters stand on their own and can be read individually.